

EE4146 Data Engineering and Learning Systems

Test 1 Solution and Marking Scheme

Semester B, 2023-24

The test consists of 3 parts.

PART B Part B consists of 3 questions, carrying 28 marks in total.

B.1. The following two 4-dimensional vectors data are collected from 2 types of topics, including stories and fictions. **(10 marks)**

Stories	Fictions
[0.6, 0.1, 2.0, -1.0]	[1.9, 3.6, -0.4, -2.0]

We are also given the below 3 different documents 4-d vectors from unknown topics:

D1	[0.9, -0.1, 1.2, -0.3]
D2	[1.3, 1.2, -0.9, -0.3]
D3	[0.7, 0.8, 0.8, -0.9]

If the cosine similarity between a document and a topic is greater than 0.5, then we can consider this document is in the area of the topic. Use cosine similarity to classify if the above documents D1, D2, D3 are in the area of stories or fictions. Show your work to justify the answers.

Solution: Let us first calculate the similarity between the corpus data and document data.

$$\text{Cos}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{(a_1 \cdot b_1 + \dots + a_N \cdot b_N)}{(\sqrt{a_1^2 + \dots + a_N^2}) \cdot (\sqrt{b_1^2 + \dots + b_N^2})}$$

$\text{Cos}(\text{Stories}, \text{D1}) = 0.90$	$\text{Cos}(\text{Fictions}, \text{D1}) = 0.22$
$\text{Cos}(\text{Stories}, \text{D2}) = -0.13$	$\text{Cos}(\text{Fictions}, \text{D2}) = 0.85$
$\text{Cos}(\text{Stories}, \text{D3}) = 0.81$	$\text{Cos}(\text{Fictions}, \text{D3}) = 0.79$

B.2. Below shows a 1-D data with 10 data points. There are 3 clusters. Find the 3 clusters using K-means. You are given that the 3 initial clusters are 18, 12 and 8. Your work should have converged in 2 iterations. Give the cluster centers (CC) in i.e., $CC1 = 18$, $CC2 = 12$, $CC3 = 8$ and cluster elements ie., $C1(a)$; $C2(b, c)$; $C3(d, e, f, g, h, j, k)$ for the 1st and 2nd iteration. **(12 marks)**

	26	18	16	14	13	12	11	9	8	4
<i>LABEL</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>j</i>	<i>k</i>

Solution and scheme:

Initial CC (cluster center) $CC1^0 = 18$, $CC2^0 = 12$, $CC3^0 = 8$

1st iteration:

Elements $C1^1$: (a, b, c) , $C2^1$: (d, e, f, g) , $C3^1$ (h, j, k)

$CC1^1 = 20$; $CC2^1 = 12.5$; $CC3^1 = 7$ (let them off if the numerical value is not very very accurate)

2nd and final iteration:

Elements $C1^1$: $(a, b,)$, $C2^1$: (c, d, e, f, g) , $C3^1$ (h, j, k)

$2\% \times 3 = 6\%$

$CC1^2 = 22$; $CC2^2 = 13.2$; $CC3^2 = 7$

$2\% \times 3 = 6\%$

B.3. Below is a table showing the proximity matrix of 5 universities. Use hierarchical clustering to find the tree diagram of these 5 universities. Draw the tress diagram as your answer and give the key values (merged values of different cells) in the tree diagram. **(6 marks)**

Solution: Step 1: The given proximity matrix.

	CityU	Birmingham	Hull	UCL	Durham
CityU	0	59.4	63.9	118.3	106.9
Birmingham		0	5.7	69	60.1
Hull			0	68.5	60.6
UCL				0	12.2
Durham					0

The nearest pair is (Bir/Hull) 5.7, merge them to one cluster.

Step 2: Update the proximity matrix.

	CityU	UCL	Durham	Bir/Hull
CityU	0	118.3	106.9	59.4
UCL		0	12.2	68.5
Durham			0	60.1
Bir/Hull				0

The nearest pair is (UCL/Durham) 12.2, merge them to one cluster.

Step 3: Update the proximity matrix.

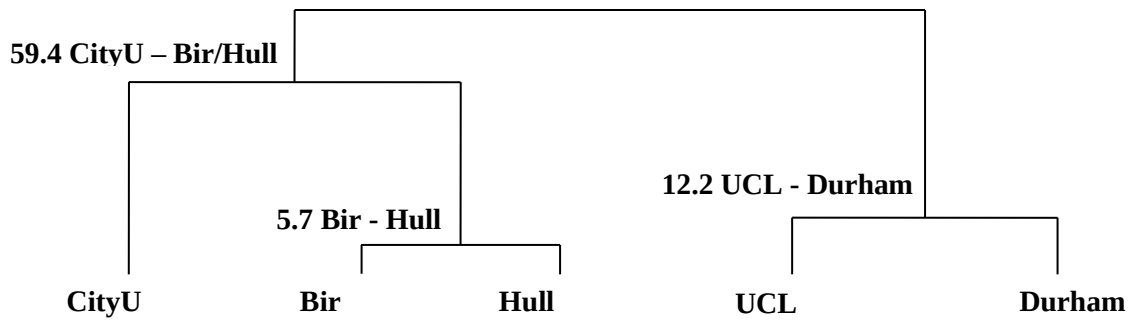
	CityU	Bir/Hull	UCL/Durham
CityU	0	59.4	106.9
Bir/Hull		0	60.1
UCL/Durham			0

The nearest pair is (CityU/(Bir/Hull)) 59.4, merge them to one cluster.

Step 4: Update the proximity matrix.

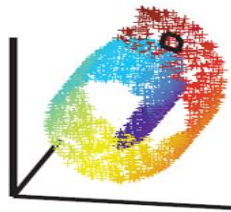
	UCL/Durham	CityU/Bir/Hull
UCL/Durham	0	60.1
CityU/Bir/Hull		0

The hierarchical tree is as the following:



PART C *Part C consists of 6 questions and each question carries 5 marks.*

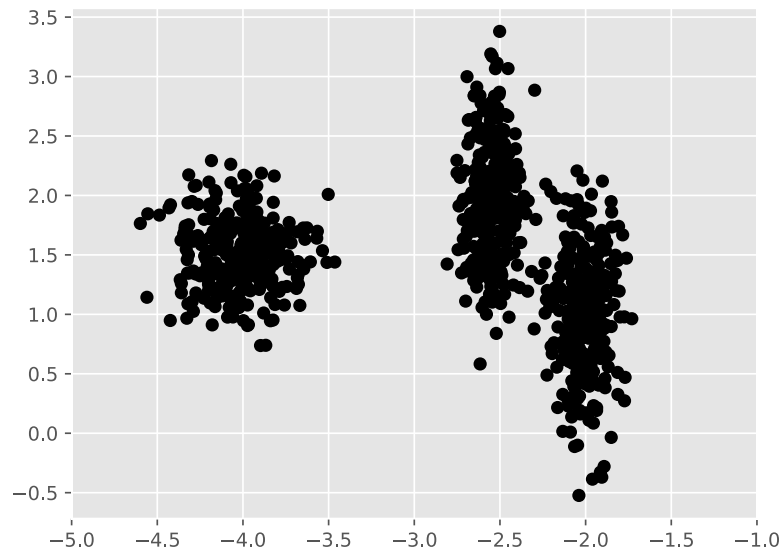
C.1. The below figure shows a dataset that we need to do dimensionality reduction from 3-D to 2-D. In this process, the result must preserve the original data information as close as possible, i.e., we need to keep the “red dots data information inside the dark ring in the reduced 2-D space. Can we use PCA to do the job? If yes, why? If no, why?



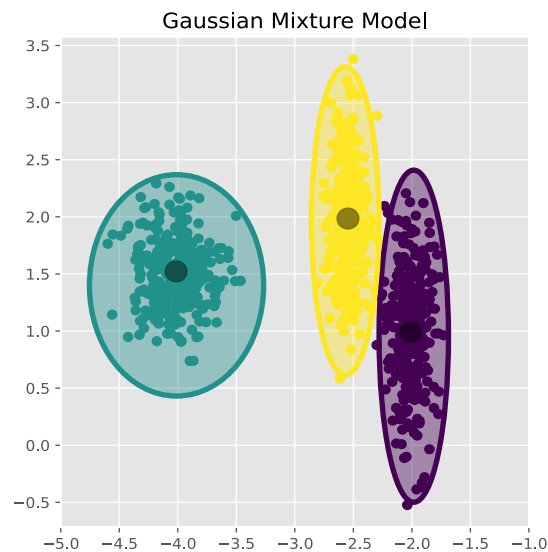
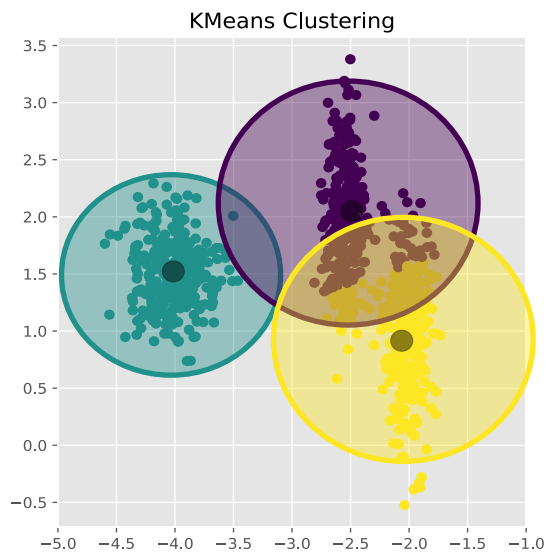
Answer will **NOT** be awarded with score if no correct justification/explanation is given.

Answer: No PCA, because PCA is linear and the red dots info inside the dark ring will be distorted or lost (or merged with the blue data info) mapping in the 2-D space.

C.2. Below are data points in a 2-D space that we need to work on its clustering. Determine if K-means or GMM clustering should be used. You MUST construct the possible clustering boundaries in the diagram to support your claim. Also use less than 10 words to justify your answer.



Answer: We must use GMM, because GMM clusters boundary is elliptical that can fit well to the dataset. K-means circle will have cluster elements being overlapped in this case.



C.3. Below table shows part of a dataset used for classifying elderly dementia disease. The dataset consists of 7 attributes with over 500 data points and a class label of “demented” or “not demented”. The researchers try to use the supervised learning method to study its characteristics, perform classification, and try to determine if a newly given patient suffers from dementia or not. List 2 major problems of the shown dataset for supervised learning.

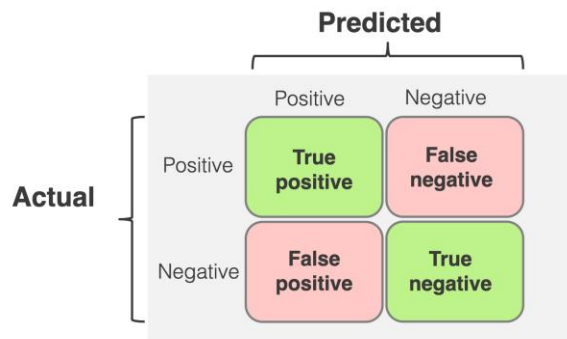
ID	Sex	Yeas of Education	Height (inches)	Weight (lbs.)	Heart Rate	Mother Demented?	Father Demented?	DEMENTED?
1	M	16	--	--	--	Yes	No	Yes
2	F	15	58	110	72	No	No	No
3	F	--	63	211	64	No	No	--
4	M	20	64	142	52	--	No	Yes
5	M	19	69	192	64	Yes	No	--
6	F	14	65	121	72	Yes	--	No
7	F	16	60	101	67	No	No	Yes
8	F	19	57	110	70	No	Yes	No
9	M	18	--	--	72	Yes	Yes	--
10	F	16	65	136	60	Yes	No	Yes

Answer: Missing link or missing data (no entry) (2%). Non-numerical data (2%) *got both correct then give 1 bonus pt to give it 5% (non-numerical: ordinal and norminal data) 1% bonus*

C.4. Name 3 different types of data and give 1 example of these 3 types of data to explain what they are.

Answer: Numerical: ie., 3, 5 *1%*, Ordinal: Very Good, Bad *2%*, Nominal: Red, Black *2%*

C.5. The confusion matrix is shown below. Define Accuracy and Precision in terms of TP (true positive), False Positive (FP), False negative (FN), and True negative (TN).



Answer: $Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$ *3%*, $Precision = \frac{TP}{TP + FP}$ *2%*

C.6 Below shows the eigen-values found in a PCA process for dealing with a 12-dimensional dataset. Determine how many dimensions and which that can be adequately used for dimensionality reduction from the view of preserving data information. You must use numerical evidence to support your answer.

Eigen-value	7.5	4	3.5	1.2	0.8	0.6	0.2	0.1	0.07	0.02	0.006	0.004
-------------	-----	---	-----	-----	-----	-----	-----	-----	------	------	-------	-------

All eigenvalues sum = 18:

First 2 eigenvalues / 18 = $11.5/18 = 63.9\%$ (not high enough)

First 3 eigenvalues / 18 = $15/18 = 83.3\%$ (that is very good enough way over 75%)

First 4 eigenvalues/ 18 = $16.2/18 = 90\%$ (no need to go to 90% too much)